Leonard Kleinrock



Massachusetts Institute of Technology
Research Laboratory of Electronics
Cambridge, Massachusetts

# Information Flow in Large Communication Nets
Proposal for a Ph.D. Thesis

Leonard Kleinrock

May 31, 1961

Information Flow in Large Communication Nets
Proposal for a Ph.D. Thesis

Leonard Kleinrock

## I. Statement of the Problem:

The purpose of this thesis is to investigate the problems associated with information flow in large communication nets. These problems appear to have wide application, and yet, little serious research has been conducted in this field. The nets under consideration consist of nodes, connected to each other by links. The nodes receive, sort, store, and transmit messages that enter and leave via the links. The links consist of one-way channels, with fixed capacities. Among the typical systems which fit this description are the Post Office System, telegraph systems, and satellite communication systems.

A number of interesting and important questions can be asked about this system, and it is the purpose of this research to investigate the answers to some of these questions. A partial list of such questions might be as follows:

(1) What is the probability density distribution for the total time lapse between the initiation and reception of a message between any two nodes? In particular, what is the expected value of this distribution?

(2) Can one discuss the effective channel capacity between any two nodes?

(3) Is it possible to predict the transient behavior and recovery time of the net under sudden changes in the traffic statistics?

(4) How large should the storage capacity be at each node?

(5) In what way does one arrive at a routing doctrine for incoming messages in different nets? In fact, can one state some bounds on the optimum performance of the net, independent of the routing doctrine (under some constraint on the set of allowable doctrines)?

(6) Under what conditions does the net jam up, i.e., present an excessive delay in transmitting messages through the net? The solution to this problem will dictate the extent to which the capacity of each link can be used (i.e., the ratio of rate to channel capacity, which is commonly known as the utilization factor).

(7) What are the effects of such things as additional intra-node delays, and priority messages?

One other variable in the system is the amount of information that each node has about the state of the system (i.e., how long the queues are in each other node). It is clear that these are critical questions which need answers, and it is the intent of this research to answer some of them.

In attempting the solution of some of these problems, it may well be that the study of a specific system or application will expose the basis for an understanding of the problem. It is anticipated that such a study, as well as a simulation of the system on a digital computer, will be undertaken in the course of this research.

## II. History of the Problem

The application of Probability Theory to problems of telephone traffic represents the earliest area of investigation related to the present communication network problem. The first work in this direction dates back to 1907 and 1908 when F. Johannsen [1][1,2] published two essays, the one dealing with delays to incoming calls in a manual telephone exchange, and the other being an investigation as to how often subscribers with one or more lines are reported "busy." It was Dr. Johannsen who encouraged A. K. Erlang to investigate problems of this nature. Erlang, an engineer with the Copenhagen telephone exchange, made a number of major contributions to the theory of telephone traffic, all of which are translated and reported in [1]; his first paper (on the Poissonian distribution of incoming calls) appeared in 1909 and the paper containing the results of his main work was published in 1917, in which

---

[1] Numerals in square brackets refer to the bibliography.

[2] Reference to Johannsen's work will be found in [1], page 16.

he considered the effect of fluctuations of service demands on the utilization of the automatic equipment in the telephone exchange.

A few other workers made some contributions in this direction around this time, and a good account of the existing theories up to 1920 is given by O'Dell [2,3]; his principal work on grading appeared in 1927. Molina[4,5] was among the writers of that time, many of whom were concerned mainly with attempts at proving or disproving Erlang's formulas, as well as to modify these formulas.

The theory of stochastic processes was developed after Erlang's work. In fact it was Erlang who first introduced the concept of statistical equilibrium, and called attention to the study of distributions of holding times and of incoming calls. Much of modern queuing theory is devoted to the extension of these basic principles with the help of more recent mathematical tools.

In 1928, T.C. Fry [6] published his book (which has since become a classic text) in which he offered a fine survey of congestion problems. He was the first to unify all previous works up to that time. Another prominent writer of that period was C. Palm [7,8], who was the first to use generating functions, in studying the formulas of Erlang and O'Dell. His works appeared in 1937-1938. During this time, a large number of specific cases were investigated, using the theories already developed, in particular lost call problems. Both Fry and Palm formulated equations (now recognized as the Birth and Death equations) which provide the foundation for the modern theory of congestion.

In 1939, Feller [9] introduced the concept of the Birth and Death process, and ushered in the modern theory of congestion. His application was in physics and biology, but it was clear that the same process characterized many models useful in telephone traffic problems. Numerous applications of these equations were made by Palm [10] in 1943 In 1948, Jensen (see [1]) also used this process, without mentioning its name, for the elucidation of Erlang's work. Kosten [11], in 1949, studied the probability of loss by means of generalized Birth

and Death equations. Waiting line and trunking problems were explained by Feller [12] in his widely used book on probability, making use of the theory of stochastic processes. At around 1939, the problems of waiting lines and trunking problems in telephone systems were taken up more by mathematicians than by telephone engineers.

In 1950, C.E. Shannon [13] considered the problem of storage requirements in telephone exchanges, and concluded that a bound can be placed on the size of such storage, by estimating the amount of information used in making the required connections. In 1951, F. Riordan [14] investigated a new method of approach suitable for general stochastic processes. R. Syski [15], in 1960, published a fine book in which he presented a summary of the theory of congestion and stochastic processes in telephone systems, and also cast some of the more advanced mathematical descriptions in common engineering terms.

In the early 1950's, it became obvious that many of the results obtained in the field of telephony were applicable in much more general situations, and so started investigations into waiting lines of many kinds, which has developed into modern Queuing Theory, itself a branch of Operations Research. A great deal of effort has been spent on single node facilities, i.e., a system in which "customers" enter, join a queue, eventually obtain "service" and upon completion of this service, leave the system. P.M. Morse [16] presents a fine introduction to such facilities in which he defines terms, indicates applications, and outlines some of the analytic aspects of the theory. P. Burke [17], in 1956, showed that for independent inter-arrival times (i.e., Poisson arrival), and exponential distribution of service times, the inter-departure times would also be independent (Poisson). In 1959, F. Foster [18] presented a duality principle in which he shows that reversing the roles of input (arrivals) and output (service completions) for a system will define a dual system very much like the original system. In contrast to the abundant supply of papers on single node facilities, relatively few works have been published on multi-node facilities (which is the area of interest to this thesis). Among those papers which have been presented

is one by G.C. Hunt [19] in which he considers sequential arrays of waiting lines. He presents a table which gives the maximum utilization factor (ratio of average arrival rate to maximum service rate) for which steady state probabilities of queue length exist, under various allowabal queue lengths between various numbers of sequential service facilities. J.R. Jackson [20], in 1957, published a paper in which he investigated networks of waiting lines. His network consisted of a number of service facilities into which customers entered both from external sources as well as after having completed service in another facility. He proves a theorem which stated roughly, says that a steady state distribution for the system state exists, as long as the effective utilization factor for each facility is less than one, and in fact this distribution takes on a form identical to the solution for the single node case.[3] In 1960, R. Prosser [21] offered an approximate analysis of a random routing procedure in a communication net in which he shows that such procedures are highly inefficient but extremely stable (i.e., they degrade gracefully under partial failure of the network).

The two important characteristics of the communication nets that form the subject of this thesis are (1) the number of nodes in the system is large, and (2) each node is capable of storing messages while they wait for transmission channels to become available. As has been pointed out, Queuing Theory has directed most of its effort so far, toward single node facilities with storage. There has been, in addition, a considerable investigation into multi-node nets, with no storage capabilities, mainly under the title of Linear Programming (which is really a study of linear inequalities and convex sets). This latter research considers, in effect, steady state flow in large connected nets, and has yielded some interesting results. One problem which has attracted a lot of attention is the shortest route problem (also known as the travelling salesman problem). M. Pollack and W. Wiebenson [22] have presented a review of the many solutions to this problem, among which

---

[3] Jackson's work is discussed in detail in Section IV.

are Dantzig's Simplex Method, Minty's labelling method, and the Moore-D'Esopo method. W. Jewell [23] has also considered this problem in some greater generality, and, by using the structure of the network and the principle of flow conservation, has extended an algorithm due to Ford and Fulkerson in order to solve a varied group of flow problems in an efficient manner. R. Chien [24] has given a systematic method for the realization of minimum capacity communication nets from their required terminal capacity requirements (again considering only nets with no storage capabilities); a different solution to the same problem has been obtained by Gomory and Hu [25]. In 1956, P. Elias, A. Feinstein, and C.E. Shannon [26] showed that the maximum rate of flow through a network, between any two terminals, is the minimum value among all simple cut-sets. Also, in 1956, Z. Prihar [27] presented an article in which he explored the topological properties of communication nets; for example, he showed matrix methods for finding the number of ways to travel between two nodes in a specific number of steps.

In 1959, P.A.P. Moran [28] wrote a monograph on the theory of storage. The book describes the basic probability problems that arise in the theory of storage, paying particular attention to problems of inventory, queuing, and dam storage. It represents one of the few works pertaining to a system of storage facilities.

The results from Information Theory [29] also have relation to the communication net problem considered here. Most of the work there has dealt with communication between two points, rather than communication within a network. In particular, one of the results says that there is a trade-off between message delay and probability of error in the transmitted message. Thus if delays are of no consequence, transmission with an arbitrarily low probability of error can be achieved. However, it is not obvious as yet, what effect such additional intranode delays would have in a large network of communication centers; it seems that some maximum additional delay exists, and if so, this would restrict the use of coding methods, and perhaps put a non-zero lower bound on the error probability.

## III. Discussion of Proposed Procedure

The problems associated with a multi-terminal communication net, as posed in the first section, appear to be too difficult for analysis, in an exact mathematical form. That is to say, the calculation of the joint distribution of traffic flow through a large (or even small) network is extremely difficult. Even for networks in which no feedback is present, the mathematics is unmanageable; and for those with feedback, it seems hopeless to attempt an exact solution. The question, then, is to what degree, and in what fashion can we simplify this problem?

Since it is the complicated interconnections that cause most of the trouble, one would like to isolate each node, and perform an individual analysis on it, under some boundary constraints. The node could then be represented by the results of such an analysis. In particular, it is hoped that the node representation would be sufficiently complete, by the use of perhaps two numbers, these numbers being the mean and variance of the traffic handled by the node. Thus, instead of having to derive the complicated joint distribution of the traffic in the network, one may be able to make a fairly accurate characterization by specifying two (or at most a few) parameters.

This approach is not completely naive and without justification. Consider the linear programming techniques [21-26] mentioned in the second section of this proposal. The problems handled by such techniques have a great deal in common with the communication problem at hand. Their problem is that of solving networks in which the commodity (e.g., water, people, information) <u>flows steadily</u>. A typical problem would be that of finding the set of solutions (commonly referred to as feasible solutions) which would support a given traffic flow in a network. A solution would consist of specifying the flow capacity for each link between all pairs of nodes. In general, a large number of solutions exist, and a lot of effort has been spent in minimizing the total capacity used for such a problem. One obvious requirement is that the average traffic entering any node must be less than the total capac-

ity leaving the node. Notice that the important statistic here is the <u>average traffic flow</u>, and if the flow is <u>steady</u>, then we have a deterministic problem. Now, in what way does this problem differ from the problem considered in this proposal? Clearly, the difference is that we do not have a steady flow of traffic. Rather, our traffic comes in spurts, according to some probability distribution. Consequently, we must be prepared to waste some of our channel capacity, i.e., the channel will sometimes be idle. A good measure of how non-steady our traffic is, is the variance of the traffic. That is, for zero variance, we are reduced to the special case above, namely, steady flow. As the variance goes up, we can say less and less about the arrival time, and the traffic becomes considerably more random in time. Thus, it is reasonable to expect that the two important parameters which characterize our traffic are the mean and variance of the flow. Notice that a necessary, but clearly not sufficient condition for a feasible solution to our problem is that the average traffic entering the node must be less than the total capacity of channels leaving the node. In 1951, Kendall [30] showed for a single node with Poisson inter-arrival times (at a rate $\lambda$ per sec), an infinite allowable queue, and an arbitrary service distribution (with mean $1/\mu$ and variance $v$), that the expected waiting time in the system, $E(t)$ was

$$E(t) = (1/\mu) + \frac{(1/\mu)^2 + v}{2[(1/\lambda) - (1/\mu)]}$$

This clearly shows a linear dependence on the variance.

Reference has already been made to J.R. Jackson [20]. The assumptions that he made in his analysis of networks of waiting lines was that the arriving traffic at each node had a Poisson distribution, that the service time was exponentially distributed, and that infinite queues at each node were allowed. With these assumptions, he was able to derive the distribution of traffic at all the nodes. It is important to analyze his assumptions carefully. The Poisson assumption effectively

characterized the traffic with two parameters. This same assumption also effectively decoupled the nodes from each other. His results state that if the mean traffic satisfies the necessary condition stated in the previous paragraph, then the resultant traffic can be characterized by a two parameter description.

The question of queuing discipline is an interesting one, and one which causes some difficulties. That is, the node must decide on a method for choosing some member in the queue to be served next. An interesting simplication tothis question, and perhaps a key to the solution of the network problem may be obtained as follows. Consider that class of queue disciplines which require that a channel facility never remain idle, as long as the queue is non-empty (clearly this is a reasonable constraint). Now, adopting a macroscopic viewpoint, (i.e., removing all labels from the members in the queue), what can be said about the mean and variance of the waiting time distribution for this class of queue disciplines? It seems that some reasonably tight bounds might exist for this distribution, independent of the particular discipline used. Perhaps some other restriction on the class of disciplines will be required in order to obtain meaningful results. However, under such a set of restrictions, if we can characterize the queue sufficiently well, we may then be placed in a position to obtain some overall behavior for the network. All of the queuing problems solved to date, have considered a particular queue discipline (the microscopic viewpoint), and so the results have been specialized to an extremely large degree. Adopting the macroscopic viewpoint seems to be a natural step, and it is the intention of this research to investigate this avenue.

There appear to be a number of conflicting interests in a network of this type. The things to be considered are: storage capacity at each node; channel capacity at each node; and message delay. There exists a trading relationship among these quantities, and it is necessary to attach some quantitative measures to this trade. In fact, if one wishes to generalize one step further, one can consider a multi-terminal communication system, in which the signals are perturbed by

noise in the system. Information theory tells us how to combat this disturbance, and the solution introduces additional delays in message transmission and reception. What effect these additional delays will have on the system is not clear; in fact it becomes difficult to state just what the overall capacity is for such a situation. Questions such as these are extremely important, and deserve attention.

From the statements in this section, it is clear that an approximate analysis is all that can easily be obtained for the network under consideration. Hopefully, the approximate answers will be reasonably useful. One way to check the utilization of the results is simulation. It is fully expected that, in the course of this research, a simulated net of this type will be programmed on one of the local digital computers. The author has access to the Lincoln Laboratory TX-2 computer, as well as the IBM 709. This simulation study should serve as a useful check on the results and perhaps, will also serve as a guide into the research.